Ordering cancer mutational profiles of cross-sectional copy number alterations

Ordering cancer mutational profiles of cross-sectional copy number alterations

Alex Graudenzi[1,2] Giulio Caravagna[1,3]
Iuliana M. Bocicor[3] Claudia Cava[2] Marco Antoniotti[1,5]
Giancarlo Mauri[1,6]

[1]Dept. of Informatics, Systems and Communication, University of Milan-Bicocca, Milan, Italy
[2]Institute of Molecular Bioimaging and Physiology, National Research Council (IBFM-CNR), Milan, Italy
[3]School of Informatics, University of Edinburgh, Edinburgh, UK
[4]Faculty of Mathematics and Computer Science, Babes-Bolyai University, Romania
[5]Milan Center for Neuroscience, University of Milan-Bicocca, Milan, Italy
[5]SYSBIO Centre of Systems Biology, Milano, Italy

Correspondence: *alex.graudenzi@unimib.it*

Abstract

Understanding the dynamical evolution of cancer, with the final goal of developing effective techniques for diagnosis, prediction and treatment is one of the main challenges of modern biosciences. In this paper we approach the *temporal ordering reconstruction problem*, which refers to the temporal sorting of a collection of static biological data. The solution of this problem may help in better understanding the key principles and properties of the disease progression. By using a previously proposed technique for extracting temporal progressions from cross-sectional cancer gene expression data, we develop a novel methodology to be applied to static cross-sectional copy number alterations, and we test it on patients diagnosed colorectal cancer at different stages. To capture distinct aspects of this complex phenomenon, we define several measures of chromosomal alterations and filters targeting significant portions of chromosomes. Results obtained with various measures and filters highlight the best setting for the problem, the most relevant chromosomal alterations and emphasize the influence that copy number alterations hitting key genes may have on the development of the disease.

Keywords: Colorectal cancer, Copy number alterations, Temporal ordering, Cancer progression.

1 Introduction

Cancer is a complex disease caused by different factors and events, mostly affecting the dynamics of the gene regulatory networks and of the signaling pathways that normally rule the correct functioning of cells, tissues and organs. As such, understanding and characterizing cancer dynamics and development is one of the big challenges of modern biosciences, in conjunction with the development of suitable and effective techniques for diagnosis, prediction and treatment. To this end, the collection and the in-depth analysis of experimental biological data is fundamental in order to validate theories and models, both from the qualitative and the quantitative point of view, and to foster the formulation of new hypotheses and experimental directions.

Yet, even though the amount of cancer-related data publicly available in the various databases is nowadays huge [The Cancer Genome Atlas Network (2016)], the challenge is now to deal with the dynamical characterization of such a complex phenomenon. Therefore, extracting useful information about cancer progression (i.e., a form of "dynamical" information) from "static" biological data would have a major significance and impact on the related research.

This paper approaches the "*temporal ordering reconstruction problem*" (TOR), that is the sorting of a collection of multi-dimensional biological data to reflect an accurate temporal progression of the target disease. Despite being a general statement problem, we restrict it to considering *Copy Number Alteration* (CNA) data from a set of patients at different stages of the disease and, in particular, we focus on the case of colorectal cancer progression. There are few studies that approach various forms of the TOR problem using various types of data but, to the best of our knowledge, CNA data has not been used to solve the TOR problem as defined above, i.e., to reconstruct the temporal ordering of a given set of samples [Desper et al. (1999), Pathare et al. (2009)].

Colorectal cancer. *Colorectal cancer* (CRC) is the third most common type of cancer worldwide and the second most frequent cause of cancer-related death [Jemal et al. (2010)]. Most CRCs develop through a series of distinct morphological stages that are strongly correlated with the malfunctioning of the complex signaling networks ruling the intestinal crypt dynamics and homeostasis, which is induced by the accumulation of alterations in the function of key regulatory genes and genetic instability [Vogelstein et al. (1988), Fearon & Volgestein (1990), Kinzler & Vogelstein (2002), Frank (2007), Jass (2007), Medema & Vermulen (2011), The Cancer Genome Atlas Network (2012)]. In particular, three major forms of genetic instability in CRC have been described: microsatellite instability, epigenetic changes, e.g. DNA methylation, and chromosomal instability. These happen in 13 %, 40 % and 47 % of the cases, respectively [Ashktorab et al. (2010)]. Chromosomal instability is associated with a poorer prognosis than that in patients with microsatellite instability [Walther et al. (2009)]. The chromosomal instability usually implies gains and losses of segments of chromosomes.

In the "standard progression", which covers around the 60% of the cases,

the disease progression crosses a few major phenotypic stages, i.e., adenoma, carcinoma and metastasis, through a number of intermediated phases and minor events. The samples and data that we have analyzed so far were classified by pathologists, using histological analysis, according with the standard four stages classification for CRC: the first one being the least severe and the fourth implying cancer metastasis [Reid et al. (2009)].

Copy number alterations. Chromosomal CNAs refer to regions of the DNA that have either been deleted or duplicated a certain number of times on chromosomes. These aberrations may affect the function of certain genes, modifying their expression and have been associated with susceptibility or resistance to certain diseases. In cancer, chromosomal CNAs can also lead to activation of oncogenes and inactivation of tumor suppressor genes. Mutation or deletion, generally of both copies, leads to inactivation of tumor suppressor genes, while oncogenes become active through mutation or amplification, usually of one copy [Sheffer et al. (2009)]. These CNAs can be as large as numerical anomalies of entire chromosomes, or as small as segmental amplification or deletion of less than $10\,kb$. Although a preferred order for the genetic alterations in CRC progression exists, the total accumulation of CNAs rather than their order is likely most important [Fearon & Vogelstein (1990)].

A comprehensive picture of chromosomal gains and losses during the progression from high-grade adenomas to invasive carcinomas is found in CRC tumors [Ried et al. (1996), The Cancer Genome Atlas Network (2012)]. One class of genetic alterations involves mutations of oncogenes and tumor-suppressor genes that directly control cell birth and death, such as APC, KRAS, and p53 [Shen et al. (2007)]. The chromosome 7 amplification, which is also observed in some colon adenomas, occurs at early stages of colorectal tumor progression [Bomme et al. (1994)]. During the progression from high-grade adenomas to invasive carcinomas, other specific chromosomal aberrations become common, such as gains on $8q$, $20q$, 7, 13 and deletions on $8p$, $17p$,$18q$,$15q$ and $20q$ [Ashktorab et al. (2010)].

State of the art. The TOR problem is general and can be approached from different, often complementary, perspectives. The Machine Learning and Data Mining communities pose the general problem as "parallel" to the standard *classification* one; some studies have shown that the general problem is *NP*-complete [Cohen et al. (1999), Ramakrishnan et al. (2009)].

In Computational Biology and in Bioinformatics we can distinguish at least two complementary lines of work. On one side, the TOR problem is framed as the task of re-ordering, according to some "logic", a set of *independent quantitative "observations"* of a phenomenon. This version of the TOR problem was mostly studied with cross-sectional *gene expression* data as input [Magwene et al. (2003), Gupta & Bar-Joseph (2008), Czibula et al. (2013), Guo et al. (2014)]. On the other side, the TOR problem can be framed as the problem of *inferring causal relations* among the "observables" associated to the input obser-

vations. This version of the problem was mainly studied within cancer research. In this case CNA data extracted via, e.g., *Comparative Genomic Hybridization*, is analyzed to understand which CNA is functional to cause other CNAs. The mathematical techniques used to solve this problem span from tree-based inference to Bayesian networks, and the driving logic relies on the accumulation of CNAs in a progressing cancer [Desper et al. (1999), Pathare et al. (2009), Gerstung et al. (2011), Beerenwinkel et al. (2005), Olde Loohuis et al. (2014), Ramazzotti et al. (2015)].

Clearly, these two aspects of the TOR problem are complementary: the former sorts the input observations, i.e. the samples, according to their profiles of CNA alterations, while the latter defines causal models involving the events associated to the input observations, i.e. the CNA alterations themselves. Both versions of the TOR problem are interesting, and their solutions can be used to yield, together, a better comprehension of cancer in the form of, e.g., a CNA-level classifier for profiles/accumulations.

In this preliminary work, we focus on the first of the two interpretations of the TOR problem, extending to CNAs the work on expression data by Gupta and Bar-Joseph [Gupta & Bar-Joseph (2008)]. Albeit conceptually simple, the technique is effective and promising, and is based on the reduction of the sorting problem to a well-known route-planning problem, under two biologically reasonable assumptions over the data. We shall use this technique looking solely to CNAs, rather than gene expression; in the next sections we will recall the intuitions underlying this approach.

A different approach based on minimum spanning trees and PQ-trees is introduced by Magwene *et al.* [Magwene et al. (2003)]. The minimum spanning tree algorithm is applied on a weighted, undirected graph, in which nodes are represented by multi-dimensional microarray data. These algorithms are tested on artificial data sets, as well as on time-series gene expression data sets derived from DNA microarray experiments.

Goals. Our work leverages the work of Gupta and Bar-Joseph, applied to CNA data. In our analysis, we also provide a more fine grained look on the *events* that may characterize a change in the progression profile; as it will become clear in the paper, we look at a "segmentation" of the SNP data and not only at chromosomal arms gains and losses [Daruwala et al. (2004)]. Finally, one more goal we pursue is to provide a building block in an analysis pipeline that can be used to look at *temporal reconstruction* problems that assume an already (partially) ordered dataset [Ramakrishnan et al. (2010), Antoniotti et al. (2010)]. A preliminary version of this work was presented as a poster at NETTAB 2012, Workshop on 'Integrated Bio-Search' [Bocicor et al. (2012)].

2 The Gupta and Bar-Joseph method for static expression data

Gupta and Bar-Joseph [Gupta & Bar-Joseph (2008)] formally show that, under a model of a single gene dynamics, the correct ordering of the *static expression data sets* can be recovered by solving an instance of the *Traveling Salesman Problem* (TSP).[1]

The approach by Gupta and Bar-Joseph is based on two key hypotheses:

(i) a gene driving a specific disease does not change the direction of its expression trajectory very often (i.e., its expression level either increases or decreases);

(ii) at any time point, a gene will remain at the same expression level as the previous time point with some probability.

Statement (i) means that a gene going up at a specific time point is likely to go up or remain at the same level in the next time point. Thus, despite the fact that a gene may change its expression trajectory directions multiple times, if enough data is available then the probability of this event is

$$(1-p) < \frac{1}{2}\,,$$

where p is the probability that a gene will *not* change its direction between two successive time points. This assumption is supported by some data sets where very few genes change directions more than once, i.e. most genes responding to the condition either go up and then down, or vice versa [Gasch et al. (2000), Nau et al. (2002)].

Statement (ii), which subsumes a Markov interpretation of the process, is formalized stating that a gene increases or decreases its expression level, depending on its direction, with some probability $1-q$, thus it does not change with probability q. In general, q should be small indicating that most genes do not change in time.

Setting on these premises, the random variable

$$\mathbf{Z} = \langle Z_1(t), \ldots, Z_m(t) \rangle \tag{1}$$

is defined where $Z_i(t)$ is the expression of the i-th gene at time t [Gupta & Bar-Joseph (2008)]. This variable models the expression of the set of genes one wants to consider, m in this case, assumed to be *independent*. A one-step theorem is then proved stating that, for all possible time steps $t = 1, \ldots, T$, the inequality

$$||\mathbf{Z}(t+1) - \mathbf{Z}(t)||_1 < ||\mathbf{Z}(t+2) - \mathbf{Z}(t)||_1 \tag{2}$$

[1] This is the problem of determining the *shortest possible route* that visits, exactly once, each one of a set of cities and returns to the origin city, given the list of distances between each pair of cities. Since it is a NP-hard problem, algorithms to find approximate solutions are often used [Applegate et al. (2003)].

holds with probability at least $1 - T\mathrm{e}^{-O(m(1-q))}$. Here, the L_1 metric $||\cdots||_1$ is used, i.e.

$$||\mathbf{x} - \mathbf{y}||_1 = \sum_{i=1}^{n} |x_i - y_i|$$

for $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$. Conditions for this inequality to hold are reasonable: $q < 1$, i.e., a gene has time-varying expression, and for $p \in (1/2, 1]$, i.e., a gene is more likely to maintain its expression direction in two consecutive steps. Under the very same assumptions, this one-step theorem generalizes to

$$||\mathbf{Z}(t+1) - \mathbf{Z}(t)||_1 < ||\mathbf{Z}(t+k) - \mathbf{Z}(t)||_1 \tag{3}$$

which holds with probability at least $1 - T^2\mathrm{e}^{-O(m(1-q))}$, for $k \geq 2$.

The reduction to the TSP problem proposed by Gupta and Bar-Joseph is as follows (see the Appendix for details): a graph $G = (N, E)$ is built where the set of nodes $N = \{\mathbf{p}_i \mid i = 1, \ldots, n\}$ is composed by all the n gene expression profiles input to the TOR problem, i.e. the profiles $\mathbf{p}_i$ we want to order. The graph is strongly connected, i.e. any profile can precede any other in the TOR solution, and the entry $p_{i,j}$ of the matrix Π of distances is

$$p_{i,j} = ||\mathbf{p}_i - \mathbf{p}_j||_1 \tag{4}$$

for $i, j = 1, \ldots, n$. When $n \gg \log T$ the unique TSP path reconstructs the correct ordering with high probability.

The method was applied to 50 patients affected by *glioma*, showing a good correlation with the survival times. By the TOR solution an outperforming classifier was also defined and key oncogenes identified [Gupta & Bar-Joseph (2008)].

3 Methods

To use the Gupta-Bar-Joseph method on CNA, rather than expression data, we must be sure that such data fulfill the hypotheses recalled in the previous section [Gupta & Bar-Joseph (2008)]. This is indeed the case, as outlined in some studies [Staub et al. (2006), Habermann et al. (2007), Sheffer et al. (2009), The Cancer Genome Atlas Network (2012)]. Among others, Tsafrir *et al.* showed that changes in expression level of genes are correlated to CNAs, suggesting that particular chromosomal regions are frequently gained and overexpressed (e.g., 7p, 8q, 13q, and 20q) or lost and underexpressed (e.g., 1p, 4, 5q, 8p, 14q, 15q, and 18) in primary colon tumors [Tsafrir et al. (2006)]. They also showed that these aberrations are absent in normal colon mucosa and become more frequent as the disease advances. Reid *et al.* combining gene expression and CNA data, identified more precise selected aberrations altering the expression of multiple genes involved in CRC development [Reid et al. (2009)]. Therefore, *it is safe* to assume that CNAs:

(i) span over genes that are important for CRC development;

(*ii*) do not significantly change the alteration direction (i.e. gain/loss) very often;

(*iii*) are not likely to significantly change between two successive time points.

Thus, we are assuming the same hypotheses by Gupta and Bar-Joseph, rephrased in the context of CNAs, and their approach to solve the TOR problem can then be used in this context as well.

We remark that the application of the technique here discussed to other types of genomic data regarding, e.g., *somatic mutations* or *epigenetic phenomena* (e.g., *DNA methylation*) could be sound, provided that the two key hypotheses discussed in Sect. 2 are verified. However, the verification of these hypotheses in different contexts is not straightforward. An in-depth discussion on how proceeding to set up such context-dependent verification is a very interesting problem *per se* and we will pursue it in a future work as we further develop the underlying theory.

3.1 Measuring chromosomal alterations

Every input sample S_k is a 22-dimensional vector $S_k = \langle v_1, v_2, \ldots, v_{22} \rangle$ where v_i is (a measure of) the CNAs associated to the i-th chromosome[2]. Set $\mathcal{S} = \{S_k \mid k = 1, \ldots, n\}$ is our data set with n samples. Each value v_i is a measure capturing different aspects of CNAs; testing several measures allows to understand which one performs best with respect to CRC progression, as done in other studies concerning *genetic instability* [Paris et al. (2004), Herzog et al. (2006)].

Blaveri *et al.* used the fraction of genome altered, the number of whole chromosome changes, the number of copy number transitions within a chromosome, the total number of chromosomes containing transitions, amplifications, and deletions [Blaveri et al. (2005)]. The authors found that early-stage bladder tumors differed significantly from late-stage tumors with respect to these measures. Peng *et al.* have not found significant difference in the average number of loci with CNAs between early and late stage gastric carcinoma, while amplifications were more common in advanced cancers than in early ones [Peng et al. (2003)].

Thus, given this uncertainty, we decided to test measures which consider separately deletions or amplifications, measures accounting for the actual values of the alterations, or for the number of alterations. For chromosome k, our input data is a multiset $\mathcal{C}_k = \{x_i \in \mathbb{N}\}$. We term *value* of an alteration any of the x_i's, and we term *intensity* of an alteration its distance from the expected value 2, i.e., $|x_i - 2|$, accounting that two copies of a certain segment of DNA are expected to be present. Given $\mathcal{C}_k$ we defined the measures in top of Table 1 by aggregating deletions and amplifications. Intuitively, IA_k weights how much a CNA is amplified, VA_k gives greater weight to the actual CNA value while NA_k simply counts the number of CNAs. The measures with the *avg* superscript are

[2]We do not consider the gender-linked chromosome, in order to avoid gender-related issues [Andersen et al. (2007)].

Measures aggregating deletions and amplifications ($\mathcal{C}_k$)

measure	*type*	*measure*	*type*
$\mathrm{IA}_k = \sum_{x_i \in \mathcal{C}_k} \lvert x_i - 2 \rvert$	intensity	$\mathrm{IA}_k^{\mathrm{avg}} = \mathrm{IA}_k / \mathrm{NA}_k$	Average of IA_k
$\mathrm{VA}_k = \sum_{x_i \in \mathcal{C}_k} x_i$	value	$\mathrm{VA}_k^{\mathrm{avg}} = \mathrm{VA}_k / \mathrm{NA}_k$	Average of VA_k
$\mathrm{NA}_k = \lvert \mathcal{C}_k \rvert$	number		

Measures for deletions solely ($\mathcal{D}_k$)

measure	*type*	*measure*	*type*
$\mathrm{ID}_k = \sum_{x_i \in \mathcal{D}_k} \lvert x_i - 2 \rvert$	intensity	$\mathrm{ID}_k^{\mathrm{avg}} = \mathrm{ID}_k / \mathrm{ND}_k$	average of ID_k
$\mathrm{VD}_k = \sum_{x_i \in \mathcal{D}_k} x_i$	value	$\mathrm{VD}_k^{\mathrm{avg}} = \mathrm{VD}_k / \mathrm{ND}_k$	average of VD_k
$\mathrm{ND}_k = \lvert \mathcal{D}_k \rvert$	number		

Measures for amplifications solely ($\mathcal{A}_k$)

measure	*type*	*measure*	*type*
$\mathrm{IAmp}_k = \sum_{x_i \in \mathcal{A}_k} \lvert x_i - 2 \rvert$	intensity	$\mathrm{IAmp}_k^{\mathrm{avg}} = \mathrm{IAmp}_k / \mathrm{NAmp}_k$	Average of IAmp_k
$\mathrm{VAmp}_k = \sum_{x_i \in \mathcal{A}_k} x_i$	value	$\mathrm{VAmp}_k^{\mathrm{avg}} = \mathrm{VAmp}_k / \mathrm{NAmp}_k$	Average of VAmp_k
$\mathrm{NAmp}_k = \lvert \mathcal{A}_k \rvert$	number		

Table 1: **Chromosomal measures.** Measures defined over CNAs: top panel contains those aggregating deletions and amplifications, mid and bottom panels separate deletions from amplifications.

simply the averaged analogous. For a set $\mathcal{C}_5 = \{3, 4, 3, 3, 3, 1\}$ we have $\mathrm{IA}_5 = 7$, $\mathrm{VA}_k = 17$ and $\mathrm{NA}_k = 6$, so $\mathrm{IA}_k^{\mathrm{avg}} = 7/6$ and $\mathrm{VA}_k^{\mathrm{avg}} = 17/6$.

We also considered measures separating amplifications from deletions, as done in previous studies [Herzog et al. (2006), Paris et al. (2004)]. A deletion happens when the CNA value is $x_i < 2$, an amplification when $x_i > 2$; thus we define $\mathcal{D}_k = \{x_i < 2 \mid x_i \in \mathcal{C}_k\}$ and $\mathcal{A}_k = \{x_i < 2 \mid x_i \in \mathcal{C}_k\}$. These two sets yield the measures in the mid and bottom panels of Table 1, which have the usual meaning.

3.2 Filtering the input data: recurrent and cancer-specific genes

So far we processed any input x_i, representing some alteration, without considering its *absolute frequency* in the data set $\mathcal{S}$, or any a priori *biological knowledge*. However, it could be that certain filters applied on $\mathcal{S}$ could lead us towards obtaining more accurate orderings. In particular, it is legitimate to argue that (i) the *most recurrent* alterations and (ii) alterations in genome regions encoding *cancer-specific genes* could be privileged to infer progression.

To define both (i) and (ii) we use information on the chromosomal regions containing CNAs. Let I_{x_i} be the interval to which alteration x_i maps to, and let $no(I_{x_i}, \mathcal{S})$ be the *number of occurrences* of I_{x_i} in data set $\mathcal{S}$. Notice that

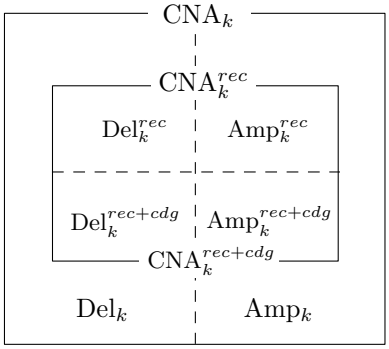

Figure 1: Input data. Euler diagram representing the nine different input sets and the relationships among them. $\mathcal{C}_k$ refers to CNAs for a specific chromosome k, $\mathcal{D}_k$ and $\mathcal{A}_k$ respectively refer to deletions and amplifications for chromosome k as well. Superscript indexes *rec* and *cdg* respectively refer to *recurrent* CNAs and *cancer driver gene*-related CNAs. We remark that $\mathcal{C}_k^{rec} \subset \mathcal{C}_k^{rec+cdg}$ and the corresponding holds for $\mathcal{D}_k^{rec}$ and $\mathcal{A}_k^{rec}$.

two alterations could belong to two (or more) partially overlapped intervals. As $no(I_{x_i}, \mathcal{S})$ increases that the DNA in I_{x_i} contains CNAs in a larger number of samples. Thus, for (i) it suffices to restrict $\mathcal{C}_k$ (or, analogously, $\mathcal{D}_k$ or $\mathcal{A}_k$) to

$$\mathcal{C}_k^{rec} = \{x_i \in \mathcal{C}_k \mid no(I_{x_i}, \mathcal{S}) > \delta_{min}\}\,.$$

For (ii) we consider both recurrent CNAs and CNAs which belong to intervals of the DNA coding for the genes presented in Table 2 [3]. So, if $\mathcal{G}$ is the set of all DNA intervals containing at least one cancer driver gene, then:

$$\mathcal{C}_k^{rec+cdg} = \mathcal{C}_k^{rec} \cup \{x_i \mid I_{x_i} \in \mathcal{G}\}\,.$$

The chromosome measures defined in Table 1 can then be applied on these filtered data sets. The data sets we discussed are graphically represented in Figure 1.

4 Results

This section presents a performance evaluation and a discussion of the results obtained by our technique. Tests were extensively performed on synthetic data, real applications on two groups of patients diagnosed with CRC.

[3] These genes are listed in the "Atlas of Genetics and Cytogenetics in Oncology and Hematology", a database containing oncogenes, cytogenetics data, clinical entities in cancer and cancer-prone diseases [Huret et al. (2001)].

Gene	Chr.	Roles in promoting CRC
APC	5q21-22	Wnt is (de)activated to degrade the b-catenin oncoprotein.
P53	17p13	Responsible for cell-cycle arrest and a cell-death checkpoint.
MLH1 MSH2 MSH6	3p21.3 2p22-p21 2p16	Mutations associated with defective mismatch repair.
PMS2	7p22.2	Post-replication DNA mispairs correction.
AXIN2	17q23-q24	Regulation of the beta-catenin stability in the Wnt pathway.
STK11	19p13.3	Encodes a serine threonine kinase and regulates cell polarity.
PTEN	10q23.3	PI3K activation yielding cell-survival and apoptosis suppression.
BMPR1A	10q22.3	Involved in the BMP/TGF-beta pathway.
SMAD4	18q21.1	Mediates the TGFβ pathway suppressing epithelial cell growth.
MYH	1p34.1	Involved in oxidative DNA damage repair.
DCC	18q21.3	Netrin-1 receptor promoting apoptosis when netrin is low.
KRAS	12p12.1	Unchecked activity of signaling through MAPK and PI3K.

Table 2: **Colorectal cancer driver genes.** Driver genes which can be considered to select events relevant to colorectal cancer progression [Huret et al. (2001), Markowitz & Bertagnolli (2009)].

4.1 Synthetic data

We evaluated the performance of our algorithm on a large set of synthetic data with various size and affected by the presence of *noise*. Each data set is generated by the following procedure, which we shortly report here and discuss in detail in the Appendix.

We hypothesize the existence of N disjoint life expectancy classes, and create M samples "naturally" clustered for each class. A sample is a vector in which every component represents the overall CNAs of a specific chromosome; its value is randomly assigned according to the values common to the sample class. The basic assumption that to a larger CNA magnitude corresponds lower survival as a consequence of the disease progression, is fulfilled by defining a total ordering among classes. We make this synthetic data more realistic by including a Binomial model of experimental error and intrinsic biological variability. This is obtained in the form of a *noise parameter* $p \in [0, 1]$. With probability p any CNA is assigned a random value, regardless of the class the sample belongs to (e.g., if $p = 0.1$ around the 10% CNAs in each sample are purely random). The goal of this analysis is to determine, as a function of p, the percentage of samples that are misclassified by the method.

In Figure 2 we display the average and best performance of the algorithm for different values of noise, number of chromosomes and number of samples for each class, with respect to the L_1 metric.[4] As expected, the method correctly orders the whole set of samples (error approx. 0, on average) when data is "clean" (i.e., $p = 0$), and dramatically drops in performance for $p \geq 0.3$. All in all, this suggest that the method might fail when intense noise is expected,

[4]Results obtained with the Euclidean distance are generally worst than those obtained with L_1, for low noise values, and equivalent for high noise values (not shown).

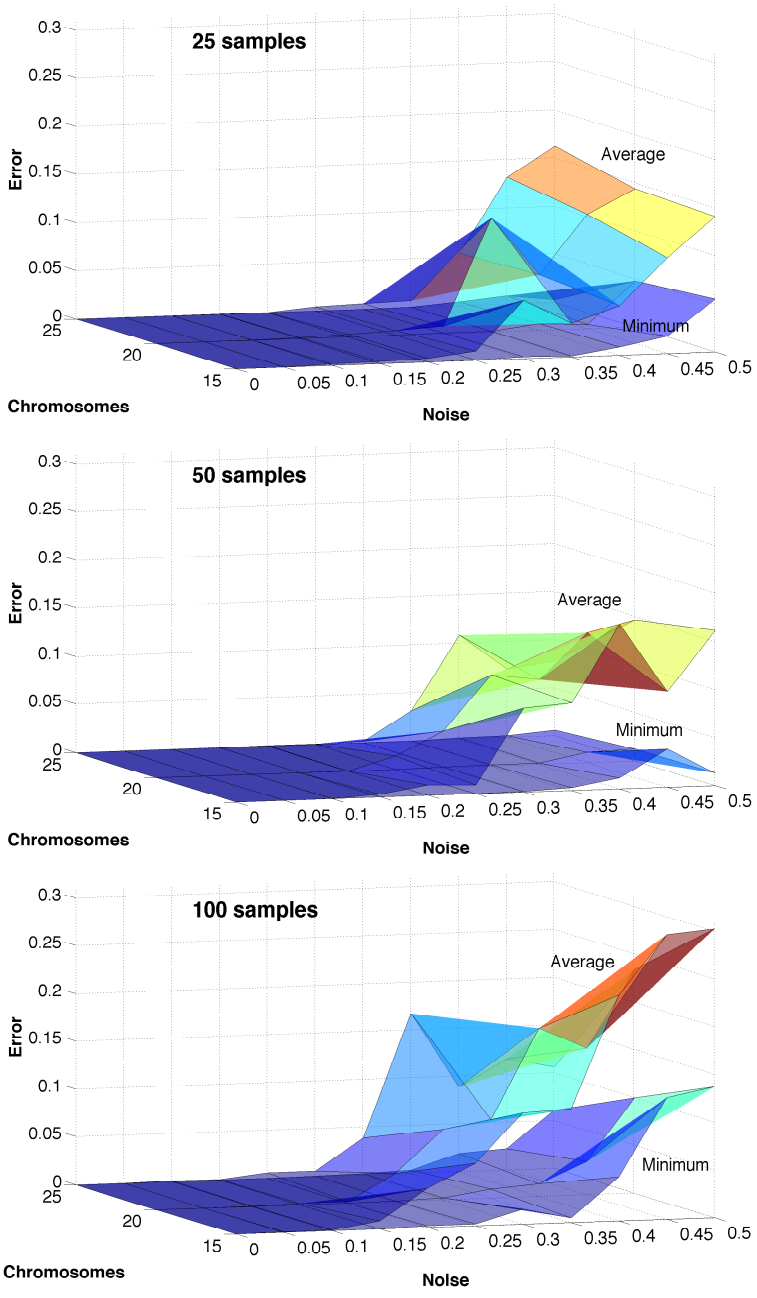

Figure 2: **Performance on synthetic data.** Average/minimum error ratio for the reconstruction method. Noise is discretized and 100 synthetic data sets with 10 life expectancy classes containing 25, 50, and 100 distinct samples are created. Combinations of 15, 20 or 25 chromosomes are used.

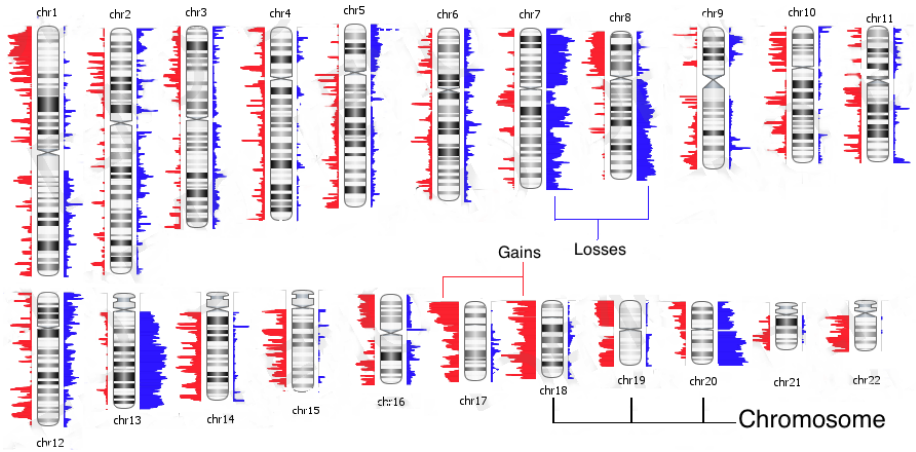

Figure 3: **Colorectal cancer - dataset 1.** Copy number alterations for dataset, divided by chromosome [Reid et al. (2009)]. Amplifications (or gains) are represented in blue, while deletions (or losses) are represented in red.

which hints at the intrinsic difficulty of inferring an effective ordering in the presence of noise.

4.2 Colorectal cancer - dataset 1

The first real life data set we used was taken from the study of Reid *et al.* [Reid et al. (2009)]. It contains tissue specimens from 53 consecutive sporadic CRCs, which we reduce to 44 since for 9 samples the stage of the tumor was unknown. The samples were hybridized to Affimetrix GeneChipVR Human Mapping 250K NspI (SNP arrays). Raw intensity CEL files of the SNP arrays were processed with CNAG program v.2.0 to detect chromosomal CNAs, using an unpaired reference of 44 HapMap normal samples [Nannya et al. (2005), Reid et al. (2009)]. With this procedure we obtain a CNA value for each, per chromosome. The CNAs for the whole set of samples are illustrated in Figure 3.

For each sample we are given a series of information, the most important being the overall survival (i.e., the survival time following the moment when samples were taken), the stage and the mutation status of certain oncogenes or tumor suppressor genes, such as APC, KRAS and p53 [Shen et al. (2007)]. These are summarized in Table 3. Notice that this data set contains only one sample classified in stage I, which however is statistically significant to our data.[5]

Three types of tests were made, one on this whole data and two when data

[5]To detect whether this sample is an outlier and if it may perturb our analyses, we performed Dixon's Q test on two sets containing the number and the values of all alterations, respectively. By a comparison we detected that this sample is not any less statistically relevant than the others [Verma & Quiroz-Ruiz (2006)].

#	Stage	Surv.	Vogelgram
1	IV	6	A-K-T-C
2	II	102+	A-K-T-C
3	IV	12	A-K-T-C
4	II	97+	A-K
5	II	53+	T-C
6	IV	26	-
7	IV	4	A-K-T-C
8	IV	10	A-K-T
9	IV	16	T
10	IV	-	T
11	III	40	A-T-C
12	II	14+	A-T-C
13	IV	10	A-K
14	I	89+	T-C
15	IV	62+	A-K-C
16	III	6	T
17	II	61+	A-C
18	IV	2	C
19	IV	11	A-K-C
20	III	35+	A-K
21	III	1	A-C
22	IV	11	A-T

#	Stage	Surv.	Vogelgram
23	III	53	A-K-C
24	IV	10	A-K-T
25	III	64+	A-K
26	IV	9	A-C
27	III	75+	-
28	IV	2	A
29	IV	7	A-K-T-C
30	IV	22	A-K-T
31	III	83+	A-K
32	IV	6	K-T
33	III	62	A
34	IV	36	A-K-C
35	IV	10	A-K
36	IV	36	A-C
37	II	93+	A-T
38	IV	-	A-K
39	II	31+	C
40	II	18+	A-K-T-C
41	IV	6	A-K-C
42	II	99+	K
43	III	65+	A-K-T-C
44	II	38	A-T-C

Table 3: **Colorectal cancer - dataset 1.** Clinical data of the CRC data samples [Reid et al. (2009)]. The overall survival refers to months, Vogelgram refers to the mutation of driver genes APC (A), KRAS (K), TP53 (T) and 18q LOH (C).

were filtered, as mentioned in Section 3.2: (*i*) all CNAs, (*ii*) recurrent CNAs and (*iii*) cancer driver genes and recurrent CNAs, which correspond to CNA_k, CNA_k^{rec} and $\text{CNA}_k^{rec+cdg}$. For each of the three types of tests mentioned above, we obtained a number of 30 different orderings: for each of the 15 types of chromosome measures that we defined two orderings were obtained by using, in turns, the L_1 and Euclidean distances. CNAs were considered recurrent when present in more than 5% of samples ($\delta_{min} = 5$) and 10%, for cases (*ii*) and (*iii*).

Results and discussion. The 90 orderings that we obtained are ranked according to the squared deviation distance from the ideal ordering, which is the one yielded by the decreasing rate of survival predicted (see the Appendix) [Reid et al. (2009)].

In Figure 4 the box plots of the *minimum* distances, for all three types of tests (*i*), (*ii*) and (*iii*), both for the L_1 and Euclidean distances (a table containing all the distances can be found in the appendix in Figure 6). It can be observed that, while the distribution of the distance are similar, with median values in the range $0.4 - 0.45$, the lowest distance, i.e. the best ordering, is

obtained (a) when CRC driver genes and recurrent CNAs are considered, (b) when the average of the alteration values is considered and (c) when Euclidean distance is used to build the TSP instance (see the Table in Fig. 6). This clearly outlines the importance of combining biological knowledge with mathematical techniques to achieve significant results, as one might expect.

The ordering is plotted in Figure 4 where its correlation with the survival time is shown. Samples in the left half of the ordering indicatively belong to patients with higher survival times, samples in the right half have lower survival times. The average survival time of the rightmost half is 22 months, while of the left half is 49. The data set contains two samples without survival time that were marked by a black asterisk. The inconsistencies between the stage and survival time are marked by either a red or a blue asterisk, respectively indicating too high/too low survival time prediction. As mentioned above, the algorithm determines the order without determining its initial sample, therefore the solution is time-reversible and the starting sample is chosen according to the biological knowledge [Gupta & Bar-Joseph (2008)]. All in all, we may observe that for the best ordering there are considerably more samples in stages III and IV in the right half (20 samples) than in the left one (13 samples). Also, the number of samples in stages I and II in the left half (9 samples) surpasses the one in the right half (2 samples). Thus, we can state that the order is, to a certain degree, also compatible with the histological stage. We also remark that the sample having stage I is positioned in the first place.

We tested different chromosomal measures: the best and second best orderings display rather similar distances (0.189 and 0.247), while the other 88 orderings have quite similar distances varying in the range 0.318–0.468. Still we can conclude that, generally, amplifications or deletions considered separately induce a better ordering when *values*, *intensities* and *numbers* are considered, and regardless of the distance metrics. There are a few cases in which all alterations yielded a better ordering. Differently, when averaged values and intensities of all alterations are considered, a better ordering is obtained, for both metrics. However, in a few cases the results separately obtained for deletions or amplifications are better. Finally, on average it seems that the two distance metrics yield quite similar orderings, the differences between them being very small. Therefore, by looking at the performance of the algorithm on synthetic data (Figure 2), we can hypothesize that the overall level of noise in the data set might be either very small or very high.

4.3 Colorectal cancer - data set 2

We also tested the technique on a public CRC microarray dataset form the Gene Expression Omnibus (GEO) database (GSE11417) [Edgar et al. (2002)]. Tumor samples and paired normal tissues were hybridized to Affymetrix Mapping 50K Xba 240 arrays [Kurashina et al. (2008)]. CNAs for each sample were obtained between pairs of tumors and normal samples. Raw intensity CEL files of the SNP arrays were processed with CNAG program, as for the first dataset.

The original data set consisted of 94 samples, but due to the fact that one of

Stage	Number of samples	Stage	Number of samples
I	3	III	37
II	45	IV	8

Table 4: **Colorectal cancer - dataset 2.** Clinical data of the CRC data samples from the second dataset.

them did not have any CNAs, we excluded it from this analysis. As in the case of the previous dataset, for each sample we are given the number of alterations, per chromosome. Still, as opposed to the first dataset, here only the information regarding the histological stages for each sample is provided (Table 4). As the overall survival time is missing, to compute the best result, we used the available information, i.e. the stages. We assumed that in an ideal ordering the first samples should have the least advanced stage and that the stages increase in the ordering such that the samples categorized in stage 4 are the last ones. Thus, we computed the square deviation distance for an ordering by summing up the distances between the position of each sample in the obtained ordering and the interval where the same sample should be placed in, according to stage information.

Results and discussion. Similarly to the previous experiment, we obtained 30 different orderings. Using the SDD as described above, we evaluated each of these orderings. In Fig. 5 we show the box plots of the distances, whereas in the table in Fig. 7 in the Appendix all the distances are shown. In this case the best ordering in terms of this distance was obtained for the test taking into account (a) the *CRC driver genes and recurrent CNAs*, with (b) the chromosome measure *values of alterations* and (c) the L_1 distance. We therefore remark that, as seen previously, it is important to also look at biological information, in conjunction with mathematical methods. Figure 5 illustrates the best ordering, plotted against the histological stages. We notice that the number of samples having stages I and II is higher for the left half (28 samples) compared to the right one (20 samples). In addition, we observe the inverse tendency in what concerns the samples in stages III and IV: there are less in the left half (19 samples), compared to the right one (26 samples). Another interesting observation is that all the samples having stage IV are correctly placed in the second half.

Also in this experiment all alterations proved to be more important than just deletions or amplifications. The only difference is that here the values of alterations lead to the best ordering, while before it was the average of these values. As respects the two used distances, we remark that the minimum SDDs are obtained for the L_1 distance, but we must mention that on average, the values of the SDDs obtained by using the L_1 distance are only slightly dissimilar from those recovered using the Euclidean distance (at most $5 \cdot 10^{-1}$ units).

5 Conclusions and Further Work

In this paper we have presented a particular solution for the temporal ordering reconstruction problem (TOR) as defined in Section 1; we have built our approach on a previously proposed solution [Gupta & Bar-Joseph (2008)], by adapting it to chromosomal copy number alterations (CNAs) data and we tested it on two colorectal cancer (CRC) data set. To the best of our knowledge, our work is the first to adapt the TSP approach to the TOR problem, in conjunction with CNA data; other approaches using CNAs have been published, but in their respect, we have actually implemented a finer control over the genome regions where CNAs actually appear.

In order to capture different features of the complex copy number alteration phenomenon and to detect the most relevant with respect to our objectives, we defined and used: (i) several chromosome-related measures, i.e., the intensity, the values and the number of the overall alterations, the deletions and the amplifications (and their corresponding averaged versions); (ii) distinct filters targeting significant portions of chromosomes to be considered in the analysis, i.e., recurrent CNAs and cancer driver genes-related CNAs; and (iii) different distance metrics, i.e., Euclidean distance and L_1 distance. The various combinations of these distinct criteria result in a large number of different optimal orderings, according to the specific measures, filters and metrics used.

With respect to the considered data sets, two sets of patients affected by CRC at different progression stages, the validation was achieved by using the quantitative (yet clearly partially arbitrary) measures of overall survival (or life expectancy) or the histological stages (where the overall survival was not available). The techniques based on the values (and average values) of alterations of both the recurrent and cancer-driver genes-related CNAs proved to outperform all the other techniques, pointing at a the dramatic influence that the chromosome-related alteration of some key genes has on the development of the disease (and on the related life expectancy).

Regarding the first CRC data set, we must remark that since it is somehow not completely coherent (i.e., some patients characterized by advanced stages of the CRC progression were labeled with long life expectancies and vice versa), the best ordering with respect to the overall survival cannot be optimal with respect to the histological classification as well. We believe that is most likely due, on the one hand, to the scarce number of samples in our data set and, on the other hand, to the “hidden factors" that lucidly influence the survival expectancy and that were not considered in the analysis (e.g., age, gender, etc.).

Nevertheless, this outcome does not discredit the validity, the efficiency and the general applicability of the methodology.

As for future development of the present work, the first problem that we will address in deeper detail will be the issue of noise, given the limited robustness of the method with respect to relatively noisy data, as one can see looking at Figure 2. Secondly, we plan to extend the evaluation, by proposing new chromosome-related measures. The copy number alterations are characterized by several distinct features (e.g., segment length, number, type, position) while

the measures we used are mostly univariate. Hence, extending the definitions of the several chromosome-related measures, in order to capture a wider range of characteristic properties, could lead to better and more accurate results. Last, but not least, we will investigate how combining the presented approach with classification algorithms influences the results: after the samples are classified, the algorithm could be applied in order to obtain a partial temporal ordering for each class and thus reconstruct the final ordering.

Acknowledgement

This project was partially supported by grants from the SysBioNet project, a MIUR initiative for the Italian Roadmap of European Strategy Forum on Research Infrastructures (ESFRI), by the ASTIL program, project "RetroNet", grant n. 12-4-5148000-40; U.A 053, and by NEDD Project [ID14546A Rif SAL-7] Fondo Accordi Istituzionali 2009.

References

Andersen, C. L., Wiuf, C., Kruhoffer, M., Korsgaard, M., Laurberg, S. & Orntoft, T. F. (2007), 'Frequent occurrence of uniparental disomy in colorectal cancer.', *Carcinogenesis* **28**, 38–48.

Antoniotti, M., Carreras, M., Farinaccio, A., Mauri, G., Merico, D. & Zoppis, I. (2010), 'An Application of Kernel Methods to Gene Cluster Temporal Meta-Analysis', *Computers and Operations Research* **37**(8).

Applegate, D. R., Bixby, V. & Chvátal, W. C. (2003), 'Implementing the dantzig-fulkerson-johnson algorithm for large traveling salesman problems', *Math. Programming* **97**, 91–153.

Ashktorab, H., Schäffer, A. A., Daremipouran, M., Smoot, D. T., Lee, E. & Brim, H. (2010), 'Distinct genetic alterations in colorectal cancer', *PLoS ONE* **5**(1), e8879.

Beerenwinkel, N., Rahnenführer, J., Däumer, M., Hoffmann, D., Kaiser, R., Selbig, J. & Lengauer, T. (2005), 'Learning multiple evolutionary pathways from cross-sectional data.', *Journal of computational biology : a journal of computational molecular cell biology* **12**(6), 584–598.

Blaveri, E., Brewer, J. L., Roydasgupta, R., Fridlyand, J., DeVries, S., Koppie, T., Pejavar, S., Mehta, K., Carroll, P., Simko, J. P. & Waldman, F. M. (2005), 'Bladder cancer stage and outcome by array-based comparative genomic hybridization.', *Clin Cancer Res* **11**(19 Pt 1), 7012–22.
URL: *http://www.biomedsearch.com/nih/Bladder-cancer-stage-outcome-by/16203795.html*

Bocicor, I., Caravagna, G., Graudenzi, A., Cava, C., Mauri, G. & Antoniotti, M. (2012), Ordering copy number alteration data to analyze colorectal cancer progression, *in* M. Masseroli, P. Romano & F. Lisacek, eds, 'EMBnet.journal: Proceedings of NETTAB 2012, Workshop on 'Integrated Bio-Search", Vol. 18, pp. 84–86.

Bomme, L., Bardi, G., Pandis, N., Fenger, C. & Kronborg, O. . (1994), 'Clonal karyotypic abnormalities in colorectal adenomas: clues to the early genetic events in the adenoma-carcinoma sequence.', *Genes Chromosomes Cancer* **10**(3), 190–196.

Cohen, W. W., Schapire, R. E. & Singer, Y. (1999), 'Learning to order things', *J Artif Intell Res* **10**, 243–270.

Cook, W. (2011), 'Concorde TSP solver'.
URL: *http://www.tsp.gatech.edu/concorde.html*

Czibula, G., Bocicor, I. & I.G., C. (2013), 'Temporal ordering of cancer microarray data through a reinforcement learning based approach', *PLoS ONE* **8**(10), 10.1371/annotation/21ae47e8–3ca1–41be–9262–4bc17eb445d.

Daruwala, R., Rudra, A., Ostrer, H., Lucito, R., Wigler, M. & Mishra, B. (2004), 'A versatile statistical analysis algorithm to detect genome copy number variation', *Proceedings of the National Academy of Sciences of the United States of America (PNAS)* **101**(46), 16292.

Desper, R., Jiang, F., Kallioniemi, O.-P., Moch, H., Papadimitriou, C. H. & Schaffer, A. A. (1999), 'Inferring tree models for oncogenesis from comparative genome hybridization data', *Journal of Computational Biology* **6**(1), 37–51.

Edgar, R., Domrachev, M. & Lash, A. (2002), 'Gene expression omnibus: Ncbi gene expression and hybridization array data repository', *Nucleic Acids Res.* **30**(1), 207–210.

Fearon, E. & Vogelstein, B. (1990), 'Genetic model for colorectal tumorigenesis.', *Cell* **61**, 759–767.

Fearon, E. & Volgestein, B. (1990), 'A genetic model for colorectal tumorigenesis', *Cell* **61**, 759–767.

Frank, S. (2007), *Dynamics of Cancer*, Princeton University Press,.

Gasch, A. P., Spellman, P. T., Kao, K. M., Carmel-Harel, O. & Eisen, M. B. (2000), 'Genomic expression programs in the response of yeast cells to environmental changes', *Molecular Biology of the Cell* **11**(12), 4241–4257.

Gerstung, M., Eriksson, N., Lin, J., Volgestein, B. & Beerenwinkel, N. (2011), 'The temporal order of genetic and pathway alterations in tumorigenesis', *PLoS ONE* **6**(11), 1–9.

Guo, J., Guo, H. & Wang, Z. (2014), 'Inferring the temporal order of cancer gene mutations in individual tumor samples', *PLoS ONE* **9**(5), e98676.

Gupta, A. & Bar-Joseph, Z. (2008), 'Extracting dynamics from static cancer expression data', *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **5**(2), 172–182.

Habermann, J. K., Paulsen, U., Roblick, U. J., Upender, M. B., McShane, L. M., Korn, E. L., Wangsa, D., Krüger, S., Duchrow, M., Bruch, H.-P. P., Auer, G. & Ried, T. (2007), 'Stage-specific alterations of the genome, transcriptome, and proteome during colorectal carcinogenesis.', *Genes, chromosomes & cancer* **46**(1), 10–26.
URL: *http://dx.doi.org/10.1002/gcc.20382*

Herzog, C. R., Desai, D. & Amin, S. (2006), 'Array cgh analysis reveals chromosomal aberrations in mouse lung adenocarcinomas induced by the human lung carcinogen 4-(methylnitrosamino)-1-(3-pyridyl)-1-butanone.', *Biochem Biophys Res Commun* **341**(3), 856–63.
URL: *http://www.biomedsearch.com/nih/Array-CGH-analysis-reveals-chromosomal/16455056.html*

Huret, J., Dessen, P. & Bernheim, A. (2001), 'Atlas of genetics and cytogenetics in oncology and haematology, updated', *Nucleic Acids Res.* **29**, 303–304.

Jass, J. R. (2007), 'Classification of colorectal cancer based on correlation of clinical, morphological and molecular features', *Histopathology* **50**, 113–130.

Jemal, A., Siegel, R., Xu, J. & Ward, E. (2010), 'Cancer statistics 2010', *CA Cancer J. Clin.* **60**, 277–300.

Kinzler, K. W. & Vogelstein, B. (2002), Colorectal tumors, *in* B. Vogelstein & K. W. Kinzler, eds, 'The Genetic Basis of Human Cancer (2nd edition)', McGraw-Hill, New York.

Kurashina, K., Yamashita, Y., Ueno, T., Koinuma, K., Ohashi, J., Horie, H., Miyakura, Y., Hamada, T., Haruta, H., Hatanaka, H. et al. (2008), 'Chromosome copy number analysis in screening for prognosis-related genomic regions in colorectal carcinoma', *Cancer science* **99**(9), 1835–1840.

Magwene, P. M., Lizardi, P. & Kim, J. (2003), 'Reconstructing the temporal ordering of biological samples using microarray data', *Bioinformatics* **19**(7), 842–850.

Markowitz, S. D. & Bertagnolli, M. M. (2009), 'Molecular Basis of Colorectal Cancer', *New England Journal of Medicine* **361**(25), 2449–2460.
URL: *http://www.nejm.org/doi/abs/10.1056/NEJMra0804588*

Medema, J. P. & Vermulen, L. (2011), 'Microenvironmental regulation of stem cells in intestinal homeostasis and cancer', *Nature* **474**(7351), 318–326.

Nannya, Y., Sanada, M., Nakazaki, K., Hosoya, N., Wang, L., Hangaishi, A., Kurokawa, M., Chiba, S., Bailey, D. K., Kennedy, G. C. & et al. (2005), 'A robust algorithm for copy number detection using high-density oligonucleotide single nucleotide polymorphism genotyping arrays', *Cancer Research* **65**(14), 6071–6079.
URL: *http://www.ncbi.nlm.nih.gov/pubmed/16024607*

Nau, G. J., Richmond, J. F., Schlesinger, A., Jennings, E., Lander, E. S. & Young, R. A. (2002), 'Human macrophage activation programs induced by bacterial pathogens', *Proc Nat'l Academy of Sciences* **99**, 1503–1508.

Olde Loohuis, L., Caravagna, G., Graudenzi, A., Ramazzotti, D., Mauri, G., Antoniotti, M. & Mishra, B. (2014), 'Inferring tree causal models of cancer progression with probability raising', *PLoS ONE* **9**(12), e115570.

Paris, P. L., Andaya, A., Fridlyand, J., Jain, A. N., Weinberg, V., Kowbel, D., Brebner, J. H., Simko, J., Watson, J. E. V., Volik, S., Albertson, D. G., Pinkel, D., Alers, J. C., van der Kwast, T. H., Vissers, K. J., Schroder, F. H., Wildhagen, M. F., Febbo, P. G., Chinnaiyan, A. M., Pienta, K. J., Carroll, P. R., Rubin, M. A., Collins, C. & van Dekken, H. (2004), 'Whole genome scanning identifies genotypes associated with recurrence and metastasis in prostate tumors.', *Hum Mol Genet* **13**(13), 1303–13.
URL: *http://www.biomedsearch.com/nih/Whole-genome-scanning-identifies-genotypes/15138198.html*

Pathare, S., Schaffer, A., Beerenwinkel, N. & Mahimkar, M. (2009), 'Construction of oncogenetic tree models reveals multiple pathways of oral cancer progression', *Int J Cancer* **124**(12), 2864–2871.

Peng, D.-F., Sugihara, H., Mukaisho, K.-i., Tsubosa, Y. & Hattori, T. (2003), 'Alterations of chromosomal copy number during progression of diffuse-type gastric carcinomas: metaphase- and array-based comparative genomic hybridization analyses of multiple samples from individual tumours.', *J Pathol* **201**(3), 439–50.
URL: *http://www.biomedsearch.com/nih/Alterations-chromosomal-copy-number-during/14595756.html*

Ramakrishnan, N., Patnaik, D. & Sreedharan, V. (2009), 'Temporal Process Discovery in Many Guises', *Computer* **42**(8), 97–101.

Ramakrishnan, N., Tadepalli, S., Watson, L. T., Helm, R. F., Antoniotti, M. & Mishra, B. (2010), 'Reverse engineering dynamic temporal models of biological processes and their relationships.', *PNAS* **107**(28), 12511–12516.

Ramazzotti, D., Caravagna, G., Olde Loohuis, L., Graudenzi, A., Korsunsky, I., Mauri, G., Antoniotti, M. & Mishra, B. (2015), 'CAPRI: efficient inference of cancer progression models from cross-sectional data', *Bioinformatics* **31**(18), 3016–3026.

Reid, J. F., Gariboldi, M., Sokolova, V., Capoblanco, P., Lampis, A., Perrone, F., Signoroni, S., Costa, A., Leo, E., Pilotti, s. & Pierotti, M. A. (2009), 'Integrative approach for prioritizing cancer genes in sporadic colon cancer', *Genes, Chromosomes and Cancer* **48**, 953–962.

Ried, T., Knutzen, R., Steinbeck, R., Blegen, H., Schröck, E., Heselmeyer, K., Du Manoir, S. & Auer, G. (1996), 'Comparative genomic hybridization reveals a specific pattern of chromosomal gains and losses during the genesis of colorectal tumors.', *Genes chromosomes cancer* **15**(4), 234–245.
URL: *http://www.ncbi.nlm.nih.gov/pubmed/8703849*

Sevaux, M. & Sorensen, K. (2005), Permutation distance measures for memetic algorithms with population management, *in* 'Proceedings of the The Sixth Metaheuristics International Conference', MIC'05.

Sheffer, M., Bacolod, M. D., Zuk, O., Giardina, S. F., Pincas, H., Barany, F., Paty, P. B., Gerald, W. L., Notterman, D. A. & Domany, E. (2009), 'Association of survival and disease progression with chromosomal instability: a genomic exploration of colorectal cancer.', *Proceedings of the National Academy of Sciences of the United States of America* **106**(17), 7131–7136.
URL: *http://dx.doi.org/10.1073/pnas.0902232106*

Shen, L., Toyota, M., Kondo, Y., Lin, E., Zhang, L., Guo, Y., Hernandez, N. S., Chen, X., Ahmed, S., Konishi, K., Hamilton, S. R. & Issa, J.-P. J. (2007), 'Integrated genetic and epigenetic analysis identifies three different subclasses of colon cancer', *Proceedings of the National Academy of Sciences* **104**(47), 18654–18659.
URL: *http://dx.doi.org/10.1073/pnas.0704652104*

Staub, E., Groene, J., Mennerich, D., Roepcke, S., Klaman, I., Hinzmann, B., Castanos-Velez, E., Mann, B., Pilarsky, C., Brummendorf, T., Weber, B., Buhr, H.-J. & Rosenthal, A. (2006), 'A genome-wide map of aberrantly expressed chromosomal islands in colorectal cancer', *Molecular Cancer* **5**, 37.

The Cancer Genome Atlas Network (2012), 'Comprehensive molecular characterization of human colon and rectal cancer', *Nature* **487**(7407), 330–337.

The Cancer Genome Atlas Network (2016), 'The cancer genome atlas', Web site at `http://cancergenome.nih.gov/`.

Tsafrir, D., Bacolod, M., Selvanayagam, Z., Tsafrir, I., Shia, J., Zeng, Z., Liu, H., Krier, C., Stengel, R. F., Barany, F., Gerald, W. L., Paty, P. B., Domany, E. & Notterman, D. A. (2006), 'Relationship of Gene Expression and Chromosomal Abnormalities in Colorectal Cancer', *Cancer Res* **66**(4), 2129–2137.
URL: *http://dx.doi.org/10.1158/0008-5472.CAN-05-2569*

Verma, S. & Quiroz-Ruiz, A. (2006), 'Critical values for six dixon tests for outliers in normal samples up to sizes 100, and applications in science and engineering', **23**(2), 133–161.

Vogelstein, B., Fearon, E., Hamilton, S., Kern, S., Preisinger, A., Leppert, M., Nakamura, Y., White, R., Smits, A. & Bos, J. (1988), 'Genetic alterations during colorectal-tumor development.', *N. Engl. J. Med.* **319**, 3526–3535.

Walther, A., Johnstone, E., Swanton, C., Midgley, R., Tomlinson, I. & Kerr, D. (2009), 'Genetic prognostic and predictive markers in colorectal cancer', *Nature Reviews Cancer* **9**(7), 489–499.
URL: *http://dx.doi.org/10.1038/nrc2645*

A Appendix

A.1 Solving a TSP instance

The standard formulation of the TSP problem, requires a *TSP-solver* to retrieve, from a starting city, the shortest path visiting each city exactly once and then returning to the original city. The route is determined by a graph-like representation of the cities and the available routes. In graph theory this is equivalent to finding the *Hamiltonian cycle* with the least weight, given a complete weighted graph.

Here cities are represented by the samples, and a *distance matrix* is used to define the distances between any two samples. Two types of metrics are used to determine an overall distance between two vector samples: the L_1 *distance* $||\cdot||_1$ introduced in Section 2

$$||\mathbf{x}-\mathbf{y}||_1 = \sum_{i=1}^{n} |x_i - y_i|\,,$$

and the *Euclidean distance* $||\cdot||$

$$||\mathbf{x}-\mathbf{y}|| = \sum_{i=1}^{n} \sqrt{(x_i - y_i)^2}$$

with $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$. With n samples, given that each chromosomal measure and both the Euclidean or the L_1 distances are independent of n, building such a matrix has a time complexity of $\Theta(n^2)$.

As shown in Section 2, the work that we used as a starting point for our methodology introduces theorems stating that the unique TSP path reconstructs the correct ordering with high probability, theorems that are proved for the L_1 distance [Gupta & Bar-Joseph (2008)]. Still, the Euclidean distance is widely used to describe (dis)similarities between genetic data, therefore we chose to use it as well. We also aim at comparing the results obtained with each type of metric in order to decide whether one is more suitable than the other for CNA data.

As in both cases the obtained distance matrices are symmetric, the problem is reduced to a symmetric version of the TSP. Nonetheless, it is clear that in

our case the first and the last samples in the ordering should be distant from each other (i.e., the distance between their representing vectors should be large), considered the underlying hypotheses on copy number alterations described in the previous sections. Consequently, we solve the *Open TSP*, which searches for the shortest route that visits each city exactly once, but without returning to the origin city. The equivalent problem in graph theory is to find the *Hamiltonian path* having the minimum weight in an undirected complete weighted graph.

We used the CONCORDE tool to approximate the TSP solution [Applegate et al. (2003)]. This software for the symmetric TSP combines linear programming and cutting planes to solve the instance, and has been used to obtain the optimal solutions to the full set of 110 TSPLIB instances [Cook (2011)]. The Concorde TSP Solver [Applegate et al. (2003)] solves the classical TSP. Therefore, we convert our Open TSP into a classical one by specifying exactly the starting node for the path, and by adding a supplementary "fake" node with distance zero to the first node and very large distance in relation to all other nodes. Then, the solution returned by the Concorde TSP Solver will be the cycle beginning and ending with the "fake" node. As soon as this node is removed from the solution, we obtain a non-cyclic path, representing the temporal ordering. However, as we do not know a priori which sample should be the first in the ordering, we try all of them and in the end, we choose the ordering that has the minimum cost of the path (i.e., minimum sum of distances between samples).

The order of samples obtained by Concorde will be, with high probability, the correct temporal ordering reflecting cancer progression, retrieved by using CNA data.

A.2 Synthetic data

We hypothesize the existence of N life expectancy classes C_i, $i = 1, \ldots, N$, the first class C_1 denoting the highest life expectancy and C_N the shortest one. We then create a set of M samples S_m, $m = 1, \ldots, M$ belonging to each class. A sample is again a j-dimensional vector $V_{C_i}^{S_m}$ in which every value represents the overall CNAs of a specific chromosome and j is the number of the considered chromosomes. Every (integer) value of each vector is generated with uniform probability with respect to disjoint value ranges for the distinct classes (e.g., $[0, 5]$, $[5, 10]$, $[10, 15]$, ...), being the values in $V_{C_1}^{S_m}$, in the lower range and those in $V_{C_N}^{S_m}$, in the higher range. The assumption that a larger CNA magnitude implies a lower survival rate as a consequence of the more advanced disease progression, is therefore fulfilled. Notice that the range of values assigned to each class is arbitrary, and the goal is to design an "ideal" data set $\mathcal{P}$ in which the $N \cdot M$ samples are naturally clustered according to the N classes.

We introduce *noise* as a parameter $p \in [0, 1]$ representing the probability that to a single entry in a sample is assigned a random value among all classes, with uniform probability. This introduces, in each sample, a number of "noisy" CNAs following a Binomial distribution $\mathcal{B}(j, p)$. Thus, we build from $\mathcal{P}$ a noisy data set $\mathcal{P}_p$. For instance, if $p = 0.1$ around the 10% CNAs are random and hence incompatible with the expected life expectancy for the sample class. Here

we introduced a noise as a model of potential measurement errors and intrinsic biological variability.

We discretized p in $[0, 0.5]$ and, for each of its values, we created 100 different synthetic data sets with 10 different life expectancy classes containing 100 distinct samples. Besides, we considered data sets with three distinct numbers of chromosomes, i.e. $j = 15, 20, 25$. As performance we measure the accuracy of the obtained temporal ordering, e, as the average number of samples correctly classified in each class, and we consider equivalently ordered samples within each class. By averaging ensembles of executions of our algorithm we measure the performance variation for increasing values of p.

A.3 Evaluating ordering against predicted survival

Given that a number of different orderings is obtained for each dataset, we need a way to rank them to obtain the optimal one. We use the *survival time* (or overall survival) following the moment in which the samples were taken, under the assumption that the amount of CNAs and the progression of the disease are correlated (which intuitively results in a overall shortening of the life expectancy). Therefore, we operate as follows. We compute the "ideal ordering" by sorting all the samples for decreasing life expectancy, and we pick as ideal the ordering which is more similar to such an ordering.

Similarity is measured as the *squared deviation distance* (SDD) among the sequences [Sevaux & Sorensen (2005)]. For each sample, the difference between its actual and the expected position is computed. The SDD is the sum of the squared values of these differences, for all samples. Thus, for any two orderings π and π', the SDD is

$$\text{SDD}(\pi, \pi') = \sum_{k=1}^{n} (i - j)^2, \tag{5}$$

where $\pi(i) = \pi'(j) = k$, and $i,\ j \in \{1, \ldots, n\}$. Here $i,\ j$ are the positions of sample k in the two permutations and n is the number of samples.

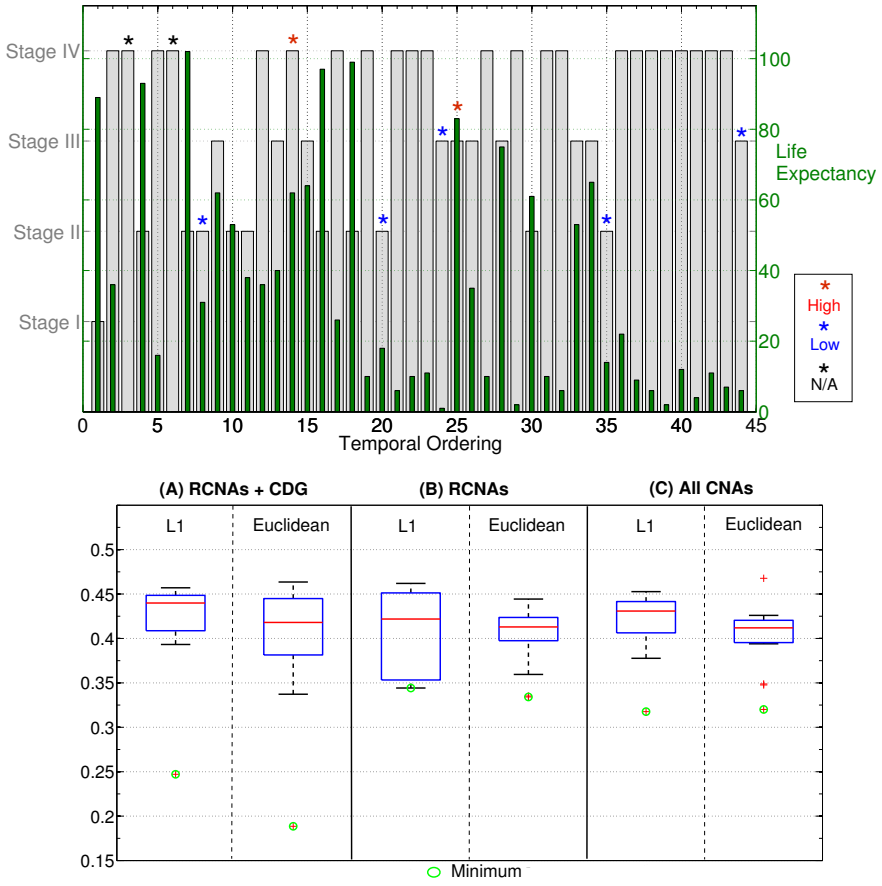

Figure 4: **Colorectal cancer - dataset 1.** Top: the best ordering predicted is plotted against both the histological stages (I–IV), and the overall survival time. Leftmost samples have higher survival times, as expected. Colored stars indicates (qualitative) mismatches between histological stage and survival time. Bottom: the box plot of the (normalized) minimum squared deviation distance (SDD) classified according to three cases: (C) all CNAs, (B) recurrent CNAs and (A) cancer driver genes and recurrent CNAs, computed with either (a) L1 or (b) Euclidean metrics. The box plots are computed in each cases on the 5 distinct chromosomal measures in Table 1. All the distances are shown in the table in Fig. 6 in the Appendix.

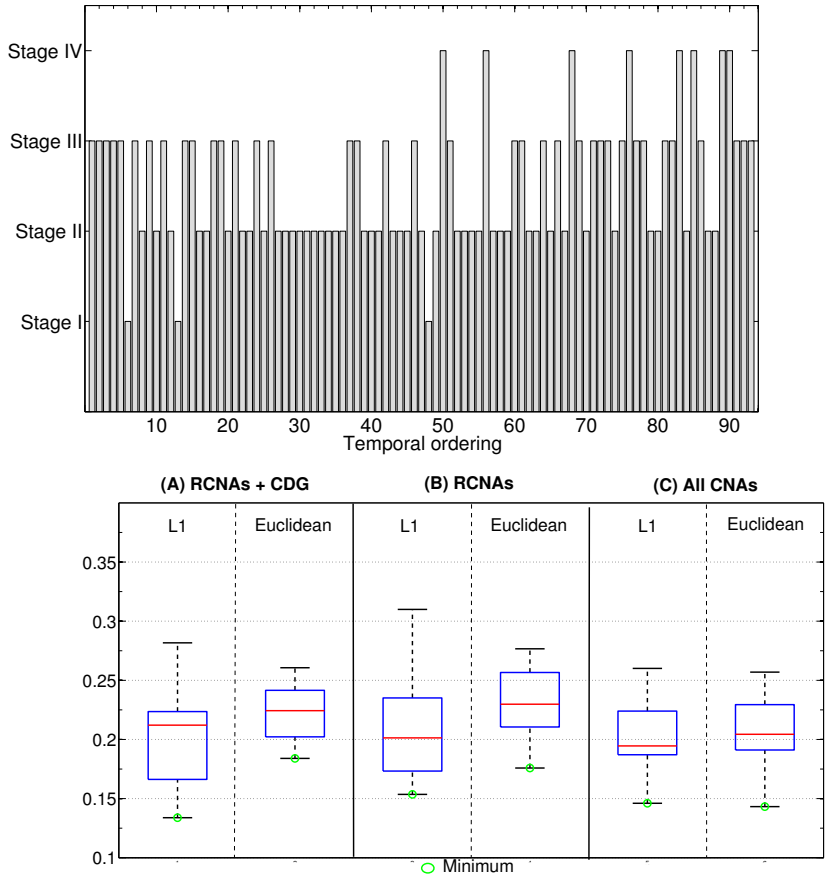

Figure 5: **Colorectal cancer - dataset 2.** Top: the best ordering predicted is plotted against both the histological stages (I–IV). No overall survival time is available for this dataset. Bottom: the box plot of the (normalized) minimum squared deviation distance SDDs classified according to three cases: (C) all CNAs, (B) recurrent CNAs and (A) cancer driver genes and recurrent CNAs, computed with either (a) L1 or (b) Euclidean metrics. The box plots are computed in each cases on the 5 distinct chromosomal measures in Table 1. All the distances are shown in the table in Fig. 6 in the Appendix.

L1									
	Simple			Recurrent			Recurrent + CDG		
	CNA	Amp	Del	CNA	Amp	Del	CNA	Amp	Del
Values	0.449	0.446	0.408	0.462	0.445	0.344	0.441	0.423	0.452
Intensity	0.318	0.439	0.404	0.422	0.436	0.345	0.440	0.457	0.425
Number	0.442	0.439	0.415	0.392	0.428	0.346	0.448	0.432	0.449
Average values	0.406	0.438	0.453	0.356	0.460	0.352	**0.247**	0.404	0.441
Average intensities	0.378	0.417	0.431	0.394	0.460	0.453	0.402	0.452	0.393

Euclidean									
	Simple			Recurrent			Recurrent + CDG		
	CNA	Amp	Del	CNA	Amp	Del	CNA	Amp	Del
Values	0.400	0.349	0.418	0.406	0.440	0.424	0.447	0.442	0.391
Intensity	0.426	0.348	0.421	0.429	0.413	0.416	0.446	0.418	0.379
Number	0.415	0.424	0.419	0.398	0.422	0.422	0.464	0.372	0.388
Average values	0.412	0.468	0.400	0.334	0.335	0.400	**0.189**	0.439	0.400
Average intensities	0.320	0.403	0.394	0.397	0.359	0.444	0.337	0.433	0.447

Figure 6: **Ranking the orderings - Dataset** 1**.** Squared deviation distance (SDD) obtained by using the different chromosome measures on the first CRC dataset. Columns refer to (A) all CNAs, (B) recurrent CNAs and (C) cancer driver genes and recurrent CNAs, divided in (1) all CNAs, (2) amplifications or (3) deletions. Rows refer to the different used measures. The values in the first table are computed with L1 distance, whereas in the second with Euclidean distance.

1.

L1									
	Simple			Recurrent			Recurrent + CDG		
	CNA	Amp	Del	CNA	Amp	Del	CNA	Amp	Del
Values	0.227	0.146	0.195	0.261	0.164	0.167	**0.134**	0.138	0.232
Intensity	0.260	0.255	0.161	0.154	0.196	0.193	0.213	0.164	0.202
Number	0.194	0.186	0.161	0.201	0.238	0.218	0.224	0.221	0.212
Average values	0.238	0.193	0.208	0.166	0.310	0.243	0.245	0.204	0.218
Average intensities	0.190	0.211	0.215	0.227	0.193	0.221	0.282	0.135	0.172

Euclidean									
	Simple			Recurrent			Recurrent + CDG		
	CNA	Amp	Del	CNA	Amp	Del	CNA	Amp	Del
Values	0.249	0.247	0.181	0.212	0.210	0.235	0.231	0.244	0.192
Intensity	0.229	0.190	0.222	0.242	0.224	0.261	0.219	0.225	0.222
Number	0.171	0.257	0.204	0.277	0.230	0.237	0.190	0.184	0.216
Average values	0.198	0.193	0.230	0.176	0.209	0.273	0.233	0.258	0.224
Average intensities	**0.143**	0.209	0.197	0.194	0.264	0.214	0.198	0.261	0.245

Figure 7: **Ranking the orderings - Dataset** 2**.** SDD obtained by using the different chromosome measures on the second CRC dataset. See the caption of Fig. 1 for a detailed description of the table headers.